Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast

Marnix H. Medema,[1,2] Eriko Takano,*[,1,3] and Rainer Breitling[2,3,4]

[1]Department of Microbial Physiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

[2]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

[3]Faculty of Life Sciences, Manchester Institute of Biotechnology, University of Manchester, Manchester, United Kingdom

[4]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

***Corresponding author:** E-mail: eriko.takano@manchester.ac.uk.

Associate editor: Katja Nowick

Abstract

The genes encoding many biomolecular systems and pathways are genomically organized in operons or gene clusters. With MultiGeneBlast, we provide a user-friendly and effective tool to perform homology searches with operons or gene clusters as basic units, instead of single genes. The contextualization offered by MultiGeneBlast allows users to get a better understanding of the function, evolutionary history, and practical applications of such genomic regions. The tool is fully equipped with applications to generate search databases from GenBank or from the user's own sequence data. Finally, an architecture search mode allows searching for gene clusters with novel configurations, by detecting genomic regions with any user-specified combination of genes. Sources, precompiled binaries, and a graphical tutorial of MultiGeneBlast are freely available from http://multigeneblast.sourceforge.net/.

Key words: **comparative genomics, gene clusters, software, evolution, operons, homology.**

Letter

Background and Rationale

Many biological systems and pathways, not only from bacteria, archaea, and fungi, but also from plants (Field and Osbourn 2008) and animals (Garcia-Fernandez 2005) are encoded by genes that are physically clustered together on the chromosome in operons or gene clusters (Fischbach and Voigt 2010). The architectures of these gene clusters are sometimes well-conserved between species, but they may also evolve quickly through rearrangements, insertions, deletions, and duplications. In many cases, knowing the evolutionary context of a gene cluster can reveal much about its function, by offering information on which other organisms possess a similar biomolecular system or pathway as encoded by the gene cluster, which parts are most strongly evolutionarily conserved, and what variants of the system or pathway exist. Homology searching can also be useful for mining large numbers of gene or operon variants from homologous gene clusters, which can then function as building blocks for the synthetic biology engineering of novel pathways or systems (Medema et al. 2012).

Although several efficient and user-friendly tools are available to perform homology searches for single genes and proteins (e.g., National Center for Biotechnology Information [NCBI]'s Basic Local Alignment Search Tool+ [BLAST+] implementation [Camacho et al. 2009]), there are few options to exhaustively mine the databases for homologs of entire operons or gene clusters. Tools such as JGI integrated microbial genomes (IMG; Mavromatis et al. 2009), PSAT (Fong et al. 2008), CCGV (Revanna et al. 2009), EDGAR (Blom et al. 2009), and Absynte (Despalins et al. 2011) each offer the possibility to perform gene neighborhood comparisons across prokaryotic genomes on precomputed data sets, but none of these allow searches against the entire GenBank database (Benson et al. 2013), nor do they allow generating custom databases from the user's own sequence data. Another tool, SynBlast (Lehmann et al. 2008), is restricted to organisms whose genetic information is deposited in ENSEMBL (Flicek et al. 2012).

Here, we present MultiGeneBlast, a comprehensive BLAST implementation to perform homology searches on multigene modules, which is built as a wrapper around NCBI BLAST+. As with the normal NCBI BLAST+ suite, the user can search the entire GenBank database or create his/her own databases. Additionally, MultiGeneBlast has the ability to perform "architecture searches," which allow finding genomic loci containing homologs of specific user-specified combinations of genes. Multiple sequence alignments of homologs can be generated automatically after the search, and all results are visualized in a user-friendly interactive eXtensible HyperText Markup Language (XHTML) page.

Implementation of the Software

MultiGeneBlast functions as a Python-based wrapper around the blastp program from the NCBI BLAST+ suite (Camacho et al. 2009), which allows detecting even distant homology between genes by using the amino acid translation as a proxy

for the gene sequence. MultiGeneBlast uses a specific database format in which each FASTA header in the database contains information on the parent nucleotide entry of the protein sequence as well as on the start and end positions and strand orientation of the gene that encodes it—besides, of course, its own functional annotation and accession number. To also make it possible to search unannotated genome sequences for homologous gene clusters, raw nucleotide databases can also be created, on which the tblastn algorithm is used instead of blastp. The MultiGeneBlast implementation (fig. 1) extends upon code written earlier for gene cluster comparison in antiSMASH (Medema, Blin, et al. 2011).

Setting up a MultiGeneBlast run can be done not only from the command line (table 1) but also with a user-friendly graphical user interface (GUI) (fig. 2) that allows easy selection of genomic regions (see our graphical tutorial in supplementary file S1, Supplementary Material online). As in our gene cluster analysis tool antiSMASH, the output is visualized in an interactive XHTML page that can be opened in a web browser. The XHTML page shows a scalable vector graphics (SVG) visualization of all sorted genomic loci (fig. 3), and clicking on a gene leads to the display of annotation information, details of any blastp/tblastn hit to the (translated) sequence of this gene (percentage identity, sequence coverage, *E*-value, and bit score), and a direct link to run an individual blastp search with the amino acid translation of this gene on the NCBI server. Optionally, multiple sequence alignments of the amino acid translations of each query gene sequence with those of its homologs can also be generated using MUSCLE (Edgar 2004).

Two Distinct Search Modes

MultiGeneBlast offers two distinct search modes: "homology search" and "architecture search." The homology search mode serves to find homologs of a known operon or gene cluster and, hence, is an extended version of a standard BLAST homology search. The input for a homology search consists of an annotated genome sequence in GBK or EMBL format, together with the start and end locations spanning the query gene cluster or operon. Alternatively to start and end sites, a list of genes can be provided that constitute the gene cluster, which has the advantage that specific genes within the gene cluster can be left out of the analysis. After running separate blastp runs for each amino acid sequence encoded in the query genomic region, MultiGeneBlast locates all hits on their parent nucleotide scaffolds in the database. Each nucleotide scaffold that received blastp/tblastn hits is then subdivided into genomic loci containing blastp/tblastn hits with a maximum mutual distance of a given number of kilobases. The default value for this distance is 20 kb, a value which has been shown to work well for most bacterial gene clusters (Medema, Blin, et al. 2011), but higher values could work better for gene clusters in fungi and plants. Similar to the ClusterBlast implementation in antiSMASH (Medema, Blin, et al. 2011), genomic loci are then sorted by an empirical similarity score $S = h + i \cdot s$, in which h represents the number of query genes with BLAST hits of at least a user-specified sequence coverage and percentage identity to the query, s represents the number of contiguous gene pairs with conserved synteny, and i represents a weighting factor that determines the weight of the synteny in determining the score. The default value for i is 0.5, which gives the number of homologous genes twice the weight as the conservation of synteny. If the obtained scores are equal, the loci are subsequently sorted by their cumulative blastp/tblastn bit scores. When testing the algorithm on a number of (semi-)manual gene cluster comparisons from the recent scientific literature, we observed that MultiGeneBlast could replicate their results accurately, as well as identify additional homologous but compositionally distinct gene clusters (supplementary file S2, Supplementary Material online).

The architecture search mode differs from a standard homology search in that the query input consists not of a known genomic region but of a FASTA file with multiple protein sequence entries, designed by the user. Thus, the user can search for all genomic loci containing a combination of certain genes within the same gene cluster. This can be of great use, for example, when searching for gene clusters encoding specific metabolic pathways containing a specified combination of enzymatic steps.

Creating Custom Databases for MultiGeneBlast

MultiGeneBlast is shipped with a database consisting of the translated amino acid sequences of all gene sequences in the GenBank database (December 12, 2012), reformatted with new FASTA headers as stated earlier. Updated versions of

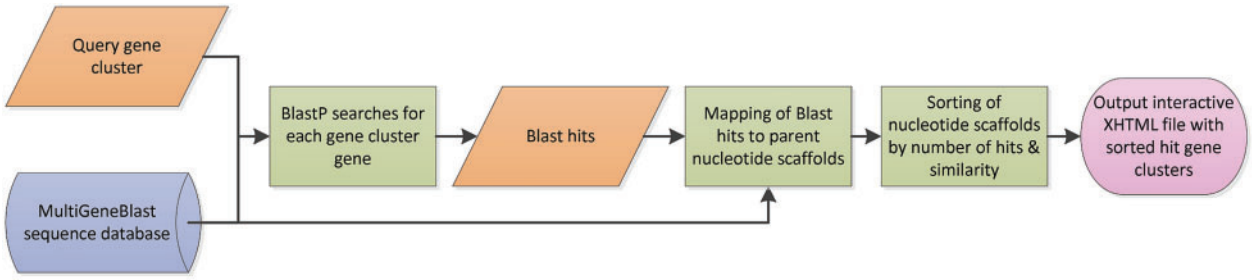

FIG. 1. Outline of the homology search process by MultiGeneBlast. First, the amino acid translation of each gene sequence within the query gene cluster is searched against the selected MultiGeneBlast database, yielding a data set of BLAST hits. The BLAST hits are then mapped to their parent nucleotide scaffolds, based on the information from the database. The nucleotide scaffolds are then sorted according to their empirical similarity scores with the query gene cluster. Finally, the sorted list of genomic loci is displayed in an interactive XHTML file that can be viewed with any modern web browser.

Table 1. Applications in the MultiGeneBlast Package.

Name	Short Description
multigeneblast	Main command-line application to run MultiGeneBlast searches
mgb_gui	GUI for configuring and starting a MultiGeneBlast run
makedb	Application to construct MultiGeneBlast databases from user data
makegbdb	Application to construct MultiGeneBlast databases from GenBank divisions
makendb	Application to construct raw nucleotide MultiGeneBlast databases from user data
makegbndb	Application to construct raw nucleotide MultiGeneBlast databases from GenBank divisions
format_embl.py	Script to generate EMBL input files from a genome sequence + gene annotations

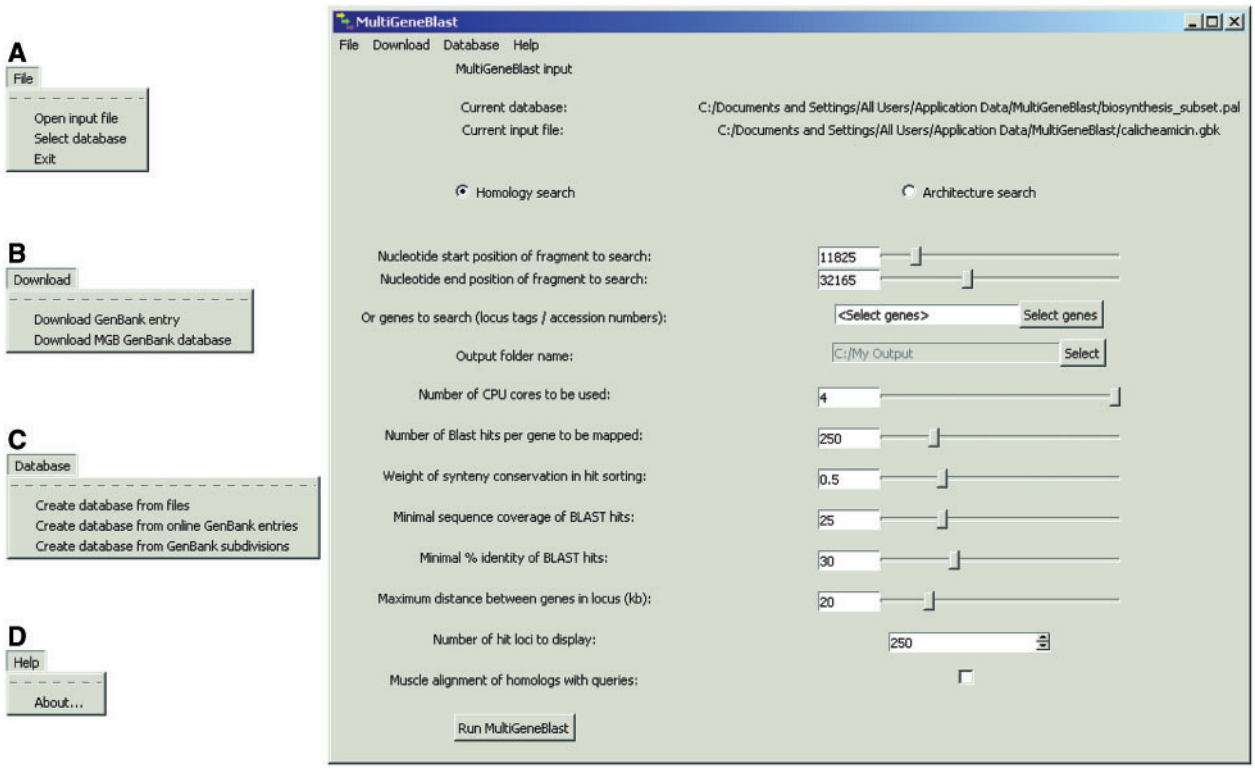

Fig. 2. A user-friendly GUI allows easy construction of databases and easy use of the program. (*A*) User-friendly selection of input files and databases. (*B*) Direct download of GenBank entries from NCBI and simple button to download MultiGeneBlast-reformatted GenBank database. (*C*) Options to design databases from files, from online GenBank entries or from entire GenBank divisions. (*D*) Link to the MultiGeneBlast website with help pages, a tutorial, and various downloads.

this database will be made available for download regularly. MultiGeneBlast also offers two tools to generate custom databases. The first tool, MakeGBDB, allows the user to construct databases from a specified subset of the GenBank subdivisions (such as BCT for bacteria and PLN for plants). The tool downloads the specified subdivisions from the NCBI FTP server and then parses them to generate a MultiGeneBlast database. The second tool, MakeDB, allows the user to construct databases from his/her own sequence data and takes as input a user-specified set of sequence files in GBK or EMBL format. For convenience, a script to generate EMBL files from nucleotide FASTA files and gene annotations is also provided.

New Approaches

MultiGeneBlast is the first full-fledged BLAST implementation that combines the input of multiple genes into a single query. Compared with previous tools for the comparative analysis of operons and gene clusters, MultiGeneBlast offers a unique set of options (table 2).

First, MultiGeneBlast allows to create databases of any combination of published and unpublished data, including the user's personal sequence data. As the costs of DNA sequencing are continuously decreasing, more and more laboratories have large amounts of unpublished sequence data that need to be analyzed before online publication. No tools have been published thus far that offer the possibility to select the user's own sequence data as both query and subject of the analysis. The IMG framework, arguably the most popular tool for gene neighborhood analysis at the moment, by design does not allow any custom queries that are outside the precomputed database, nor does it offer the option to search against custom-designed databases. In contrast, the user-friendly GUI of MultiGeneBlast makes it easy even for biologists with little or no bioinformatic expertise to design their

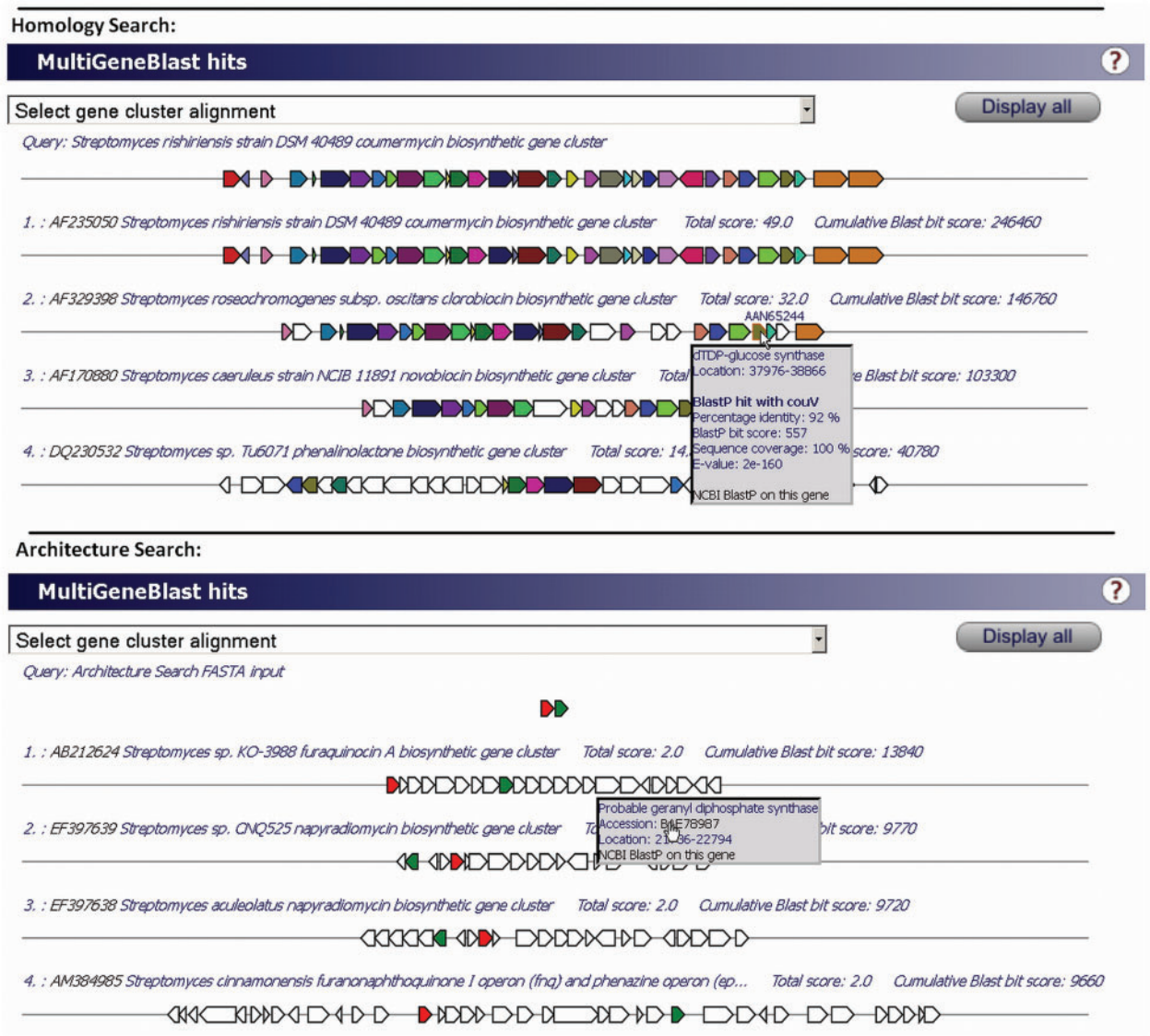

Fig. 3. Example output of a MultiGeneBlast run. The output consists of an interactive XHTML page, in which additional information on each gene appears on mouse-over or by clicking on a gene. This feature works for colored homologous genes and white nonhomologous genes. The first example output shown here displays a homology search for the coumermycin biosynthetic gene cluster, which identifies gene clusters encoding related compounds. The second example output shows the power of an architecture search to find specific pathways: By using a query of a type III polyketide synthase and a terpene cyclase, biosynthetic gene clusters encoding hybrid polyketide-terpene compounds are identified straightforwardly. Single alignments of the query gene clusters with any particular hit gene cluster can also be selected from a drop-down menu. All gene cluster images are stored in SVG format, so they can easily be transformed into publication-quality figures.

Table 2. Comparison of Different Software Tools for Gene Cluster Homology Searches.

Software	Web Tool	Stand-Alone Tool	Not Restricted to Precomputed Data	Can Search Entire GenBank Database	Based on Multiple Gene Queries	Allows Input of Personal Sequence Data	Allows Creation of Custom Databases	Architecture Search Mode	Command-Line Available	Open Source
MultiGeneBlast		X	X	X	X	X	X	X	X	X
IMG	X									
EDGAR	X									
Absynte	X					X				
PSAT	X									X
CCGV	X		X			X			X	X
SynBlast		X	X						X	X

own databases and search them with their own sequence data.

Second, most existing tools do not allow searches against the entire GenBank database but only against subsets of sequences (usually whole genome sequences) for which precomputed results have been obtained. Thousands of known and characterized gene clusters (especially biosynthetic ones) are not part of any whole-genome sequence but were instead cloned directly from the environment, or are part of a metagenomic data set, and are therefore not present in databases such as that of IMG. MultiGeneBlast, however, offers the opportunity to perform a truly exhaustive search to find all homologous genetic elements that are present in the current databases.

Third, the architecture search mode is unique to MultiGeneBlast and allows finding operons that are not similar to any operon known in advance by the user but instead contain homologs of a user-specified combination of genes.

Finally, unlike most available tools, MultiGeneBlast can be used from the command line and also generates a tab-delimited TXT output, so it can easily be integrated in a larger computational pipeline. With relatively simple scripting, large numbers of queries can thus be searched against one or more databases to perform higher-level bioinformatic analyses.

Practical Applications of MultiGeneBlast

MultiGeneBlast offers a simple and intuitive tool to perform comparative genomic analysis, facilitating functional inference and evolutionary studies of gene clusters encoding biomolecular machines or pathways.

A major application of MultiGeneBlast is to get a quick overview of the biomolecular diversity of an entire genetic element in diverse organisms and to survey all the variants that have evolved. Because MultiGeneBlast does not just display the genomic neighborhoods of one single gene but finds genomic loci with a combination of any of a list of query genes, the output will contain variants of the query genomic region consisting of any subset of that region in any arrangement. This avoids the risk of missing variants that do not contain the query gene, in contrast to approaches based on single gene input. When combining the list of identified gene cluster variants with phylogenetic information (of either species or representative genes), the evolutionary history of a gene cluster can be reconstructed, which can give valuable insight into the biomolecular functions of the various components of the encoded system. Based on patterns of evolutionary conservation, one can sometimes also get a better idea of which genes do and which genes do not belong to the gene cluster as a functional unit.

Often, distinct subclusters with separate evolutionary histories together constitute a larger gene cluster (Fischbach et al. 2008). A MultiGeneBlast analysis of the entire gene cluster may reveal its fundamental architecture, through the identification of distinct patterns of conservation of various subsets of genes from the gene cluster. This also cannot be achieved by approaches based on a single gene query.

Another important and promising application of the approach is to rapidly harvest gene parts for the synthetic biology design of biochemical pathways (Medema, Breitling et al. 2011; Medema et al. 2012). When generating synthetic versions of a particular biochemical system for heterologous implementation in a pre-engineered host, it is of great importance to test multiple versions of the system to find the one that functions best in a particular organism (Bayer et al. 2009). Because MultiGeneBlast can search the entire GenBank database, as well as any personal sequence data that may be available, it can quickly and reliably be used to identify all extant versions of an operon or gene cluster in an exhaustive manner.

Of course, many more applications of the tool are possible, as the colocalization of functionally related genes is a recurring evolutionary motif: MultiGeneBlast provides a general search tool that can be exploited in a wide range of comparative genomics studies of homologous multigene units, by expert bioinformaticians and experimental biologists alike.

Supplementary Material

Supplementary files S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank two anonymous reviewers for their constructive comments. They thank Kai Blin for contributing code originally written for the antiSMASH project. This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs (grant number STW 10463). R.B. was supported by an NWO-Vidi fellowship, and E.T. by a Rosalind Franklin Fellowship, University of Groningen.

References

Bayer TS, Widmaier DM, Temme K, Mirsky EA, Santi DV, Voigt CA. 2009. Synthesis of methyl halides from biomass using engineered microbes. *J Am Chem Soc.* 131:6508–6515.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Res.* 41:D36–D42.

Blom J, Albaum SP, Doppmeier D, Puhler A, Vorholter FJ, Zakrzewski M, Goesmann A. 2009. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* 10:154.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Despalins A, Marsit S, Oberto J. 2011. Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. *Bioinformatics* 27:2905–2906.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Field B, Osbourn AE. 2008. Metabolic diversification—independent assembly of operon-like gene clusters in different plants. *Science* 320: 543–547.

Fischbach M, Voigt CA. 2010. Prokaryotic gene clusters: a rich toolbox for synthetic biology. *Biotechnol J.* 5:1277–1296.

Fischbach MA, Walsh CT, Clardy J. 2008. The evolution of gene collectives: how natural selection drives chemical innovation. *Proc Natl Acad Sci U S A.* 105:4601–4608.

Flicek P, Amode MR, Barrell D, et al. (57 co-authors). 2012. *Nucleic Acids Res.* 40:D84–D90.

Fong C, Rohmer L, Radey M, Wasnick M, Brittnacher MJ. 2008. PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics* 9:170.

Garcia-Fernandez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet.* 6:881–892.

Lehmann J, Stadler PF, Prohaska SJ. 2008. SynBlast: assisting the analysis of conserved synteny information. *BMC Bioinformatics* 9:351.

Mavromatis K, Chu K, Ivanova N, Hooper SD, Markowitz VM, Kyrpides NC. 2009. Gene context analysis in the integrated microbial genomes (IMG) data management system. *PLoS One* 4: e7979.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* 39:W339–W346.

Medema MH, Breitling R, Bovenberg R, Takano E. 2011. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nat Rev Microbiol.* 9:131–137.

Medema MH, van Raaphorst R, Takano E, Breitling R. 2012. Computational tools for the synthetic design of biochemical pathways. *Nat Rev Microbiol.* 10:191–202.

Revanna KV, Krishnakumar V, Dong Q. 2009. A web-based software system for dynamic gene cluster comparison across multiple genomes. *Bioinformatics* 25:956–957.